PEARSON'S VERSUS SPEARMAN'S AND KENDALL'S CORRELATION COEFFICIENTS FOR CONTINUOUS DATA

by

Nian Shong Chok

BS, Winona State University, 2008

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Nian Shong Chok

It was defended on

26 May, 2010

and approved by

Thesis Advisor:
Andriy Bandos, PhD
Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Stewart Anderson, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Marike Vuga, PhD
Research Assistant Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

PEARSON'S VERSUS SPEARMAN'S AND KENDALL'S CORRELATION COEFFICIENTS FOR CONTINUOUS DATA

Nian Shong Chok, M.S.

University of Pittsburgh, 2010

The association between two variables is often of interest in data analysis and methodological research. Pearson's, Spearman's and Kendall's correlation coefficients are the most commonly used measures of monotone association, with the latter two usually suggested for non-normally distributed data. These three correlation coefficients can be represented as the differently weighted averages of the same concordance indicators. The weighting used in the Pearson's correlation coefficient could be preferable for reflecting monotone association in some types of continuous and not necessarily bivariate normal data.

In this work, I investigate the intrinsic ability of Pearson's, Spearman's and Kendall's correlation coefficients to affect the statistical power of tests for monotone association in continuous data. This investigation is important in many fields including Public Health, since it can lead to guidelines that help save health research resources by reducing the number of inconclusive studies and enabling design of powerful studies with smaller sample sizes.

The statistical power can be affected by both the structure of the employed correlation coefficient and type of a test statistic. Hence, I standardize the comparison of the intrinsic properties of the correlation coefficients by using a permutation test that is applicable to all of them. In the simulation study, I consider four types of continuous bivariate distributions composed of pairs of normal, log-normal, double exponential and t distributions. These

distributions enable modeling the scenarios with different degrees of violation of normality with respect to skewness and kurtosis.

As a result of the simulation study, I demonstrate that the Pearson's correlation coefficient could offer a substantial improvement in statistical power even for distributions with moderate skewness or excess kurtosis. Nonetheless, because of its known sensitivity to outliers, Pearson's correlation leads to a less powerful statistical test for distributions with extreme skewness or excess of kurtosis (where the datasets with outliers are more likely).

In conclusion, the results of my investigation indicate that the Pearson's correlation coefficient could have significant advantages for continuous non-normal data which does not have obvious outliers. Thus, the shape of the distribution should not be a sole reason for not using the Pearson product moment correlation coefficient.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my thesis and academic advisor, Dr. Bandos, for his encouragement, guidance, patience, time and invaluable input throughout the preparation of this work. I would also like to thank the committee members for their valuable comments and suggestions. I appreciate the feedbacks from all of them. Thank you.

Finally, I would like to thank my family and friends for their love, encouragement and support.

1.0 INTRODUCTION

In data analysis, the association of two or more variables is often of interest (e.g. the association between age and blood pressure). Researchers are often interested in whether the variables of interest are related and, if so, how strong the association is. Different measures of association are also frequent topics in methodological research.

Measures of association are not inferential statistical tests, instead, they are descriptive statistical measures that demonstrate the strength or degree of relationship between two or more variables.[19] Two variables, *X* and *Y*, are said to be associated when the value assumed by one variable affect the distribution of the other variable. *X* and *Y* are said to be independent if changes in one variable do not affect the other variable. Typically, the correlation coefficients reflect a monotone association between the variables. Correspondingly, positive correlation is said to occur when there is an increase in the values of *Y* as the values of *X* increase. Negative correlation occurs when the values of *Y* decrease as the values of *X* increase (or vise versa).[7, 15, 19]

There are many different types of correlation coefficients that reflect somewhat different aspects of a monotone association and are interpreted differently in statistical analysis. In this work, I focus on three popular indices that are often provided next to each other by standard software packages (e.g. Proc Corr, SAS v.9.2), namely the Pearson product moment correlation, Spearman's rank-order correlation and Kendall's tau correlation.

In application to continuous data, these correlation coefficients reflect the degree of association between two variables in a somewhat different manner. A strong monotonically increasing (decreasing) association between two variables usually leads to positive (negative) values of all correlation coefficients simultaneously. However, their absolute values could be quite different. Moreover for weak monotone associations, different correlation coefficients could also be of a different sign. Usually, Spearman's rank-order correlation coefficient is closer to the Pearson's than Kendall's is. However, the ordering of the true values of different correlation coefficients does not directly translated into the relative ordering of the statistical power for detecting a given type of monotone association, since the variability of the sampling distributions of different correlation coefficients could also differ substantially. Current recommendations for selecting the correlation coefficient for continuous data do not seem to incorporate statistical power considerations.

There are numerous guidelines on when to use each of these correlation coefficients. One guideline is based on the type of the data being analyzed. This guideline indicates that the Pearson product moment correlation coefficient is appropriate only for interval data while the Spearman's and Kendall's correlation coefficients could be used for either ordinal or interval data. Some guidelines also exist suggesting which correlation might be more appropriate for data that involves several types of variables.[22] According to Khamis,[11] for data that has at least one ordinal variable, Kendall's tau is more appropriate. Other investigators suggested Spearman's correlation coefficients for the same scenarios.[4, 16, 20] However, all of these correlation coefficients could be computed for interval data (e.g. continuous).[17]

The Pearson product moment correlation is a natural parameter of association for a bivariate normal distribution (it assumes zero value if and only if the two variables are

independent). Thus, a statistical test based on the Pearson's correlation coefficient is likely to be the most powerful for this type of data than similar tests on the other correlation coefficients. However, for non-normal data, the sensitivity of the Pearson product moment correlation coefficient has led to recommendations of other correlation coefficients. However, by replacing the observations by their ranks, the effect of the outliers may be reduced.[1, 2, 21] Thus, if the data contains outliers in one of both of the continuous variables, Spearman's rank-order correlation coefficient is considered more appropriate.

The aspects of conventional statistical test for the Pearson's correlation coefficient undermine it even further. Indeed, the standard procedure for testing significance of the estimates for the Pearson's correlation coefficient is sensitive to the deviations of bivariate normality.[21] Due to all these deficiencies of the Pearson's correlation coefficient, the proximity of Spearman's to Pearson's correlation coefficient in bivariate normal data,[18] and the appropriateness of Spearman's statistical test for any type of interval data makes Spearman's correlation coefficient overall more preferable.

Kendall's tau is even less sensitive to outliers and is often preferred due to its simplicity and ease of interpretation.[10] Originally, Kendall's tau correlation coefficient was proposed to be tested with the exact permutation test.[9, 10] This type of permutation test can also be applied to other types of correlation coefficient. This nonparametric procedure can help comparing the ability of the correlation coefficients to reflect a given monotone association, aside from the possible differences caused by discrepancies in the statistical testing procedures.

1.1 PEARSON PRODUCT MOMENT CORRELATION COEFFICIENT

The Pearson's correlation coefficient is a common measure of association between two continuous variables. It is defined as the ratio of the covariance of the two variables to the product of their respective standard deviations, commonly denoted by the Greek letter ρ (rho):

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

The sample correlation coefficient, *r*, can be obtaining by plugging-in the sample covariance and the sample standard deviations into the previous formula, i.e.:

$$r = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{——(1)}$$

where:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n};\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

The Pearson's correlation coefficient ranges from -1 to +1. A positive monotonic association (two variables tend to increase or decrease simultaneously) results in $\rho > 0$, and negative monotonic association (one variable tends to increase when the other decreases) results in $\rho < 0$. ρ of 0 corresponds to the absence of the monotonic association, or absence of any association in the case of bivariate normal data. However, for bivariate distributions other than bivariate normal distribution, the Pearson's correlation can be zero for dependent variables. For example, it can be '0' for the variables with non-monotonic relationship, e.g. $Y = X^2$, ($x \in (-1, 1)$). The absolute value of ρ indicates the strength of the monotonic relationship between the two variables.[2, 15, 17, 18, 19] ρ of 1 indicates a perfect linear relationship, i.e. $Y = a+bX$.

1.2 SPEARMAN'S RANK-ORDER CORRELATION COEFFICIENT

Spearman's rank-order correlation coefficient (denoted ρ_s) is a rank-based version of the Pearson's correlation coefficient. Its estimate or sample correlation coefficient (denoted r_s), can be written as follows:

$$r_s = \frac{\sum_{i=1}^{n}\Big(\big(rank(x_i) - \overline{rank(x)}\big)\big(rank(y_i) - \overline{rank(y)}\big)\Big)}{\sqrt{\sum_{i=1}^{n}\big(rank(x_i) - \overline{rank(x)}\big)^2 \sum_{i=1}^{n}\big(rank(y_i) - \overline{rank(y)}\big)^2}} \quad \text{____(2)}$$

where *rank(x_i)* and *rank(y_i)* are the ranks of the observation in the sample.

Spearman's correlation coefficient varies from -1 to +1 and the absolute value of ρ_s describes the strength of the monotonic relationship. The closer the absolute value of ρ_s to 0, the weaker is the monotonic relationship between the two variables. [3, 17] However, similar to the Pearson product moment correlation coefficient, Spearman's correlation coefficient can be 0 for variables that are related in a non-monotonic manner. At the same time, unlike the Pearson's correlation coefficient, Spearman's coefficient can be 1 not only for linearly related variables, but also for the variables that are related according to some type of non-linear but monotonic relationship.

1.3 KENDALL'S TAU CORRELATION COEFFICIENT

Similar to Spearman's rank-order correlation coefficient, Kendall's tau correlation coefficient is designed to capture the association between two ordinal (not necessarily interval) variables. Its estimate (denoted τ) can be expressed as follows:

$$\tau = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} sgn(x_i - x_j)sgn(y_i - y_j)}{n(n-1)} \quad \text{____(3)}$$

where:

$$sgn(x_i - x_j) = \begin{cases} 1 & if\ (x_i - x_j) > 0 \\ 0 & if\ (x_i - x_j) = 0 \\ -1 & if\ (x_i - x_j) < 0 \end{cases} ;\ sgn(y_i - y_j) \begin{cases} 1 & if\ (y_i - y_j) > 0 \\ 0 & if\ (y_i - y_j) = 0 \\ -1 & if\ (y_i - y_j) < 0 \end{cases}$$

This coefficient quantifies the discrepancy between the number of concordant and discordant pairs. Any two pairs of ranks (x_i, y_i) and (x_j, y_j) are said to be concordant when $x_i < x_j$ and $y_i < y_j$, or when $x_i > x_j$ and $y_i > y_j$, or when $(x_i - x_j)(y_i - y_j) > 0$. Correspondingly, any two pairs of ranks (x_i, y_i) and (x_j, y_j) are said to be discordant when $x_i < x_j$ and $y_i > y_j$, or when $x_i > x_j$ and $y_i < y_j$, or when $(x_i - x_j)(y_i - y_j) < 0$. Similar to the two previous correlation coefficients, Kendall's tau ranges from -1 to +1, with the absolute value of τ indicating the strength of the monotonic relationship between the two variable.[3, 17] However, Kendall's tau can be 1 for even a wider range of scenarios than Spearman's correlation coefficient.

1.4 MOTIVATION

The Pearson product moment correlation is the most frequently used coefficient for normal distributed data. On the other hand, nonparametric methods such as Spearman's rank-order and Kendall's tau correlation coefficients are usually suggested for non-normal data. However, although the advantages of the latter measures for categorical data are obvious, their benefits for analyzing continuous data are not that clear. Under certain formulations, all three types of correlation coefficients could be viewed as weighted averages of concordance indicators, and

Pearson's type of weighting could be conceptually preferable for continuous, but not necessarily normally distributed data.

Let:

$$P = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,;\; Q = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$x_i^* = \frac{x_i}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{x_i}{P};\; y_i^* = \frac{y_i}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{y_i}{Q}$$

$$\bar{x}^* = \frac{\bar{x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{\bar{x}}{P};\; \bar{y}^* = \frac{\bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\bar{y}}{Q}$$

Then, since:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j)$$

$$= n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j - \sum_{i=1}^{n}\sum_{j=1}^{n} x_j y_i + n\sum_{i=1}^{n} x_j y_j$$

$$= 2n\sum_{i=1}^{n} x_i y_i - 2n^2 \bar{x}\bar{y}$$

$$= 2n\left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right]$$

The Pearson correlation coefficient could be re-written as follows:

$$r = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \sum_{i=1}^{n}\left(\left(\frac{x_i}{P} - \frac{\bar{x}}{P}\right)\left(\frac{y_i}{Q} - \frac{\bar{y}}{Q}\right)\right)$$

$$= \sum_{i=1}^{n} \left((x_i^* - \bar{x}^*)(y_i^* - \bar{y}^*) \right)$$

$$= \sum_{i=1}^{n} x_i^* y_i^* - \bar{y}^* \sum_{i=1}^{n} x_i^* - \bar{x}^* \sum_{i=1}^{n} y_i^* + n\bar{x}^*\bar{y}^*$$

$$= \sum_{i=1}^{n} x_i^* y_i^* - n\bar{x}^*\bar{y}^*$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i^* - x_j^*)(y_i^* - y_j^*)$$

Equivalently, this can be written in terms of the signs and absolute values of the differences, i.e.:

$$r = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \{ sgn(x_i - x_j) sgn(y_i - y_j) \times |x_i^* - x_j^*| |y_i^* - y_j^*| \} \quad \text{____(4)}$$

This reformulation highlights the differences between the Pearson's and Kendall's correlation coefficients, the latter of which can be written as follows:

$$\tau = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} sgn(x_i - x_j) sgn(y_i - y_j)$$

Spearman's correlation coefficient could be written similar to (4) where the absolute differences between observations are replaced with the absolute differences between the corresponding ranks.

The product of sign functions $sgn(x_i\text{-}x_j)\ sgn(y_i\text{-}y_j)$ can be interpreted as a concordance indicator. It equals to 1 for concordant pairs and -1 for discordant pairs. Thus, all three correlation coefficients are proportional to the differently weighted averages of the same concordance indicators. The Pearson's correlation coefficient takes into account both the number and degree of concordances and discordances, whereas Kendall's tau correlation coefficient reflects only the numbers of concordances and discordances regardless of their

degree. Spearman's correlation is in between of the Pearson's and Kendall's, reflecting the degree of concordances and discordances on the rank scale.

Because of the variability of the observations drawn from a continuous distribution, small discordances between two close pairs of observations are quite possible even if the true measurements are concordant. For example, if a true average blood pressure of an older person is only slightly higher than that of a somewhat younger person, it is quite possible to observe a reverse order of single measurement. On the other hand, if the true averages are substantially different (in standardized units), the disagreement between the individual measurements is quite unlikely. Thus for continuous data, the degree of discordances and concordances often carries essential information about the correlation. The Pearson's correlation coefficient attempts to capture this information, while Kendall's tau correlation coefficient completely disregards it. Spearman's correlation coefficient, although, reflects the degree of concordances or discordances by using ranked observations, can equate some intrinsically small with substantially large discordances or concordances for small sample size. The ability of different correlation coefficients to reflect the degree of concordances/discordances could be translated into the relative statistical power when detecting association between two continuous variables, especially for smaller sample sizes. In this work, I focus on the relative statistical power of the test for association based on the Pearson's, Spearman's and Kendall's correlation coefficients.

2.0 SAMPLING DISTRIBUTIONS

I start my investigation by the preliminary study of the sampling distributions of the estimates for different correlation coefficients. The main objective is to assess the relative frequency of the positive estimates for the positively correlated data. This will provide an indirect indication of the relative power of statistical tests based on these correlation coefficients, and one would expect to observe a higher frequency of positive estimates for the correlation coefficient that corresponds to a more powerful statistical test. Based on the considerations presented in the previous section, I expect to observe a higher frequency of positive estimates for the Pearson's correlation coefficient, followed by Spearman's and finishing with Kendall's correlation coefficients. In addition, I compare the estimates of the true values of Spearman's and Kendall's correlation coefficients.

For this preliminary study, I compute the estimates of the correlation coefficients for 10,000 datasets generated from the bivariate normal distribution with ρ of 0.4, and sample sizes 10, 20, 50 and 100. Since the considered correlation coefficients are invariant with respect to the location-scale family of transformations, I use standard normal distributions for the two marginals, namely, $X \sim N(0,1), Y \sim N(0,1)$. Each random observation is generated as the transformations of the two independent normally distributed variable Z_x and Z_y, i.e.:

$$Z_X \sim N(0,1);\ Z_Y \sim N(0,1)$$

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Z_X \\ Z_Y \end{pmatrix} * \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Table 1 demonstrates that, as expected, the frequencies of positive estimates increase as the sample size gets larger. For sample size 100, there is no difference in the observed frequencies of positive estimates for the considered ρ of 0.4, since all estimates of the three correlation coefficients are positive. It also shows that the Pearson's correlation coefficient has the highest frequency of positive estimates, followed by Spearman's and then by Kendall's correlation coefficients. These summaries indicate that for ρ of 0.4, I could expect a reasonable magnitude of the differences in statistical power of the tests based on the considered correlation coefficients only for sample size less than 50.

Table 2 summarizes the average of the estimates of Spearman's and Kendall's correlation coefficient, that correspond to the given true value of the Pearson's correlation coefficient. The average is based on 1,000 estimated correlation coefficients each computed for sample size of 500. These results confirm the expectation that for the bivariate normal distribution, the true Pearson correlation coefficient has the largest value, followed by that of Spearman's and Kendall's. Compared with Kendall's tau correlation coefficient, the average of Spearman's rank-order correlation coefficient is closer to Pearson's. However, by itself the ordering of the true values of the correlation coefficients is not very indicative of the resulting statistical power, since the estimates can be distributed differently around these values.

Figures 1, 2 and 3 respectively summarize the sampling distributions of the Pearson's, Spearman's and Kendall's correlation coefficients for the datasets with ρ of 0.4 and sample size of 10. Each sampling distribution is obtained from 10,000 estimates. Figures 4, 5 and 6 summarize similar sampling distributions for the sample size of 20. In contrast to the Pearson's and Spearman's which have similar empirical distributions, the estimates of Kendall's tau correlation coefficient are distributed over a smaller range concentrated at around 0.26 (Table 2).

However, despite the smaller range, the frequency of positive estimates of Kendall's correlation coefficient is the lowest of the three (Table 1).

Table 1. Frequency of positive estimates of the correlation coefficients*

Sample Size	**Number of**	**Frequency**		
n	**Simulations**	**Pearson's**	**Spearman's**	**Kendall's**
10	10,000	8923	8719	8706
20		9614	9520	9473
50		9982	9963	9963
100		10,000	10,000	10,000

*The true Pearson's correlation for variables in the generated datasets is 0.4

Table 2. Estimates of the true values of different correlation coefficients

Sample Size	**Number of**	**True value**		
n	**Simulations**	**Pearson**	**Spearman's**	**Kendall's**
500	1,000	0	0	0
		0.2	0.19	0.13
		0.4	0.38	0.26
		0.6	0.58	0.41
		0.8	0.78	0.59

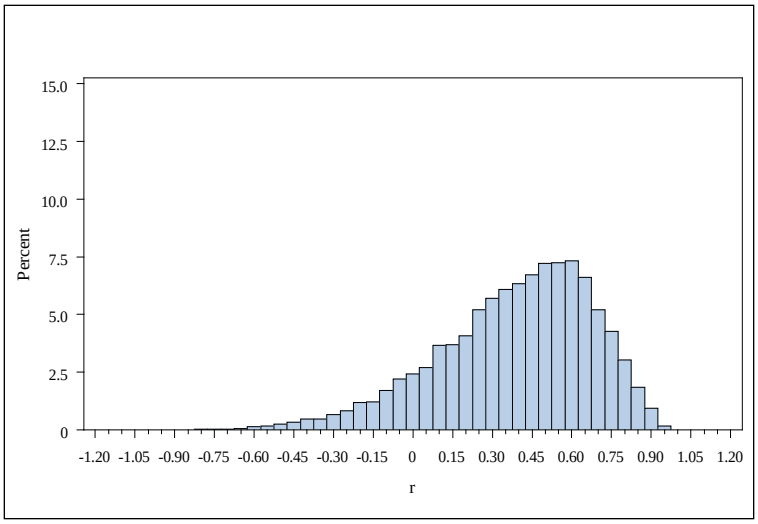

Figure 1. Histogram for the Pearson product moment correlation coefficients with n=10

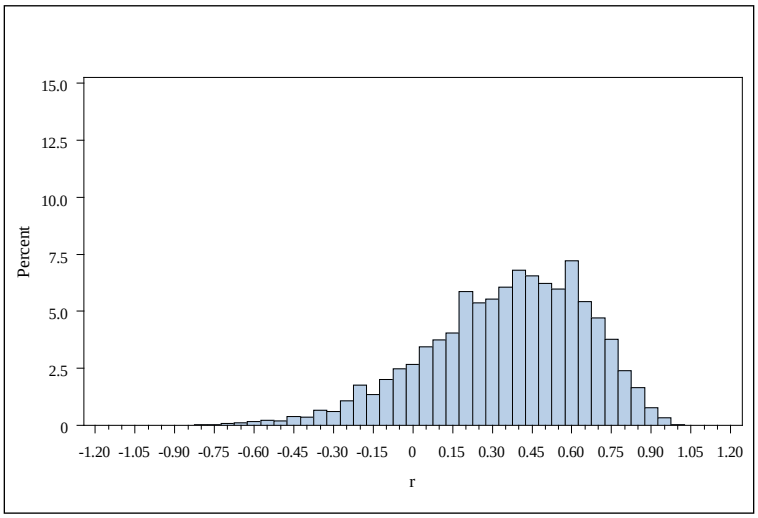

Figure 2. Histogram for Spearman's rank-order correlation coefficients with n=10

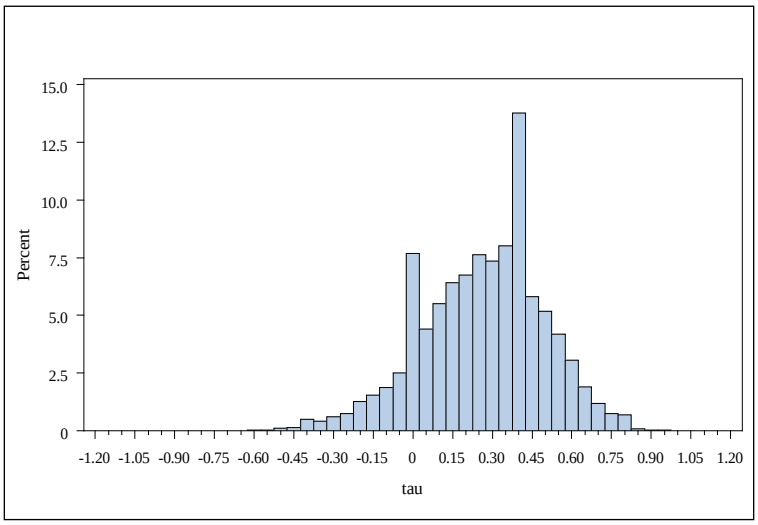

Figure 3. Histogram for Kendall's tau correlation coefficients with n=10

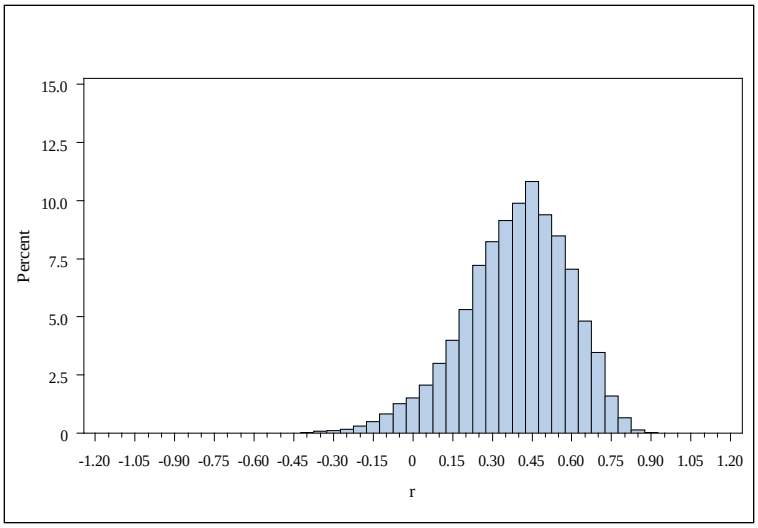

Figure 4. Histogram for the Pearson product moment correlation coefficients with n=20

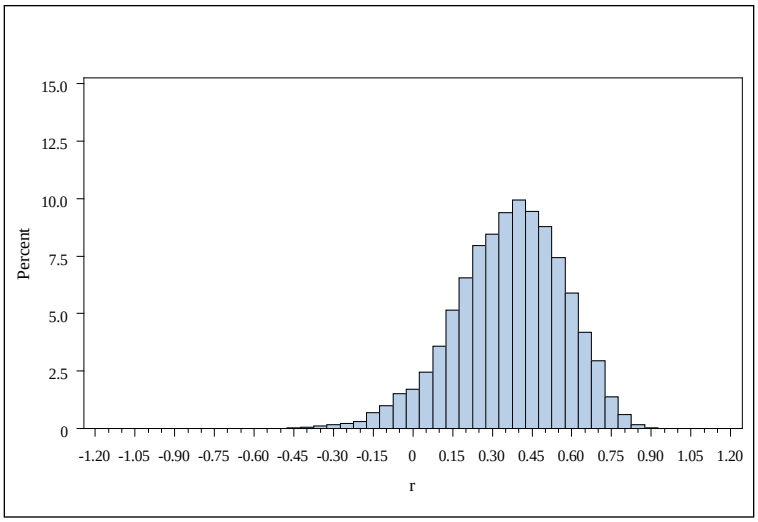

Figure 5. Histogram for Spearman's rank-order correlation coefficients with n=20

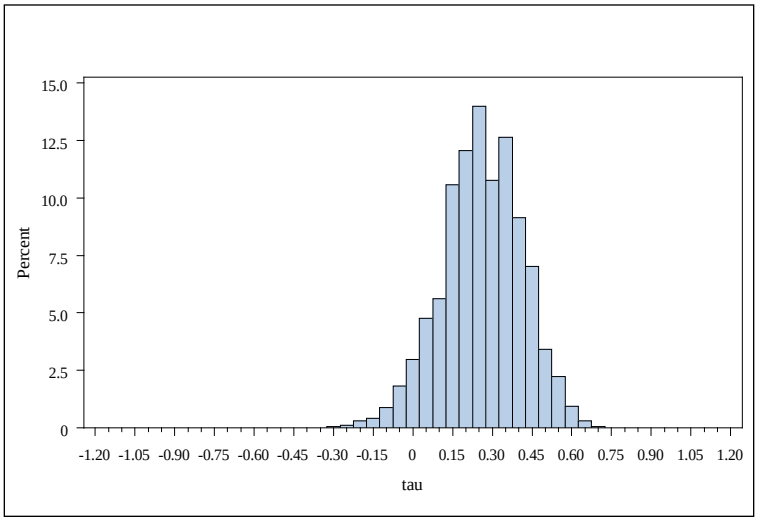

Figure 6. Histogram for Kendall's tau correlation coefficients with n=20

3.0 APPROACH

Although the preliminary results support the expectation about the relative statistical power of the tests based on different correlation coefficients, the formal conclusion require a direct investigation of the statistical power. Since the Pearson's, Spearman's and Kendall's correlation coefficients that have somewhat different methodologies for statistical analysis, I standardize my comparison of correlation coefficients by performing the same permutation test for all of them. I conduct a simulation study where permutation tests based on different correlation coefficients and estimate their type I error rate and statistical power. All programs are written in Statistical Analysis System (SAS, v.9.2) software and the code is included in Appendix.

3.1 PERMUTATION TEST

The permutation approach I consider is based on generating multiple samples from the original dataset by permuting the observations within each variable. The dataset is permuted in such a manner that the resulting permutation samples are equally likely under the null hypothesis. Under the null hypothesis of no association between the two variables, permutations of observations of one variable for the fixed order of the observations of the other are equally likely. Thus, a permutation sample can be generated by randomly re-ordering observations of

the second variable while keeping the original order for the first variable. The same permutation scheme was originally proposed for the Kendall's tau correlation coefficient.[9, 10]

By computing the values of a given statistic (here the Pearson's, Spearman's or Kendall's correlation coefficient) for every permutation sample, we can construct the permutation distributions of that statistic. The significance of the originally observed value of the statistic depends on the frequency of the observing more extreme values in the permutation distribution. I consider the mid-point p-value that quantifies the frequency of more extreme correlation coefficients plus half of the frequency of the correlation coefficients that are equal to the originally observed correlation coefficients, i.e.:

$$P = \frac{\#\ of\ (|k^*| > |k|)}{\#\ of\ permutation} + \frac{\#\ of\ (|k^*| = |k|)}{2 \times \#\ of\ permutation} \quad \text{____(5)}$$

where k^* is the value of the correlation coefficient computed from a permutation sample and k is the originally observed value of the correlation coefficient (Pearson's, Spearman's or Kendall's).

The total number of all possible permutation samples increases faster than exponentially with the increasing sample size. Under my permutation scheme, for a sample of *n* observations, there are *n!* permutation samples. Thus, instead of using all possible permutation samples, a common practice is to use a large random number of possible permutation samples. Here I base my permutation test on 5,000 of random permutations for all considered scenarios (including sample sizes from 10 to 100).

3.2 DISTRIBUTIONS USED IN THE SIMULATION STUDY

In this investigation, I consider 5 different bivariate distributions. One is the bivariate normal distribution that represents the scenario where the Pearson's correlation coefficient is expected to be better than any other measures of association. The other four distributions represent different degrees of violations of binormality. Two bivariate log-normal distributions represent the cases of distributions with moderately (γ_1 = 0.95) and severely (γ_1 = 1.75) skewed marginals. The bivariate t and double exponential distributions illustrate the cases of distributions which marginals have moderate (γ_2 = 1) and severe (γ_2 = 3) excess of kurtosis correspondingly, where excess of kurtosis is defined as the forth moment minus 3 (3 is the kurtosis for the standard normal distribution). All 5 types of distributions are generated as appropriate transformation from a standard bivariate normal distribution with a given covariance structure.

Table 3. Summary of distributions used in the simulation study

Bivariate Distribution	$\mu_x = \mu_y$*	$\sigma_x = \sigma_y$*	*df*	λ	Skewnees	Excess of Kurtosis
Normal	0	1	-	-	0	0
Log-normal	0	0.3	-	-	0.95	4.6
Log-normal	0	0.5	-	-	1.75	8.9
T	0	1	10	-	0	1.0
Double Exponential	0	1	-	1	0	3.0

* Parameters of the initial normal distributions (before transformations)

The bivariate distribution $(X^*, Y^*) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ is generated according to the following transformation approach:

$$\begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} + \sqrt{\Sigma \begin{pmatrix} X \\ Y \end{pmatrix}} \quad \text{——(6)}$$

where:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

and ρ is the correlation between x and y. Due to the invariance of the considered correlation coefficients to location-scale type of transformation for the data and without the loss of generality, I use $\mu = 0$ and $\sigma = 1$.[3, 23]

The log-normal distribution is generated as an exponential transformation of the bivariate normal distribution, i.e.:

$$e^{(X^*, Y^*)} = \left(X_{log-n}, Y_{log-n}\right) \sim Log - N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

The skewness of the resulting log-normal distribution could be computed as follows: [8]

$$\gamma_1 = \left(e^{\sigma^2} + 2\right)\sqrt{e^{\sigma^2} - 1}$$

Thus, in order to impose moderate and severe skewness, I start with the bivariate normal distributions with variances of 0.3 and 0.5 respectively. For the bivariate normal distribution with correlation ρ, the correlation for the corresponding log-normal distribution (ρ_{ln}) can be computed according to the following formula:[12]

$$\rho_{ln} = \frac{exp(\rho\sigma_x\sigma_y) - 1}{\sqrt{\{exp(\sigma_x^2) - 1\}\{exp(\sigma_y^2) - 1\}}}$$

For the considered simulation scenarios (see the next section), the values of the correlation for the log-normal distribution are approximately the same as for the original normal distribution (Table 6).

I generate the bivariate t distribution $(X_t, Y_t) \sim t(\rho, df)$ as follows:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} x^* \\ y^* \end{pmatrix} * \sqrt{df / \begin{pmatrix} u \\ v \end{pmatrix}}$$

where the independent variables u and v are generated from the gamma distribution: $(U,V) \sim 2 * GAM\left(\frac{df}{2}, 2\right)$ and (X^*, Y^*) are generated from bivariate normal scenario (6). The correlation in the original bivariate normal distribution and the resulting t distribution are related the same. Indeed:[14]

$$cov\,(X_t, Y_t) = \frac{v}{v-2}\,\rho$$

$$V(X_t) = \frac{v}{v-2};\; V(Y_t) = \frac{v}{v-2}$$

hence:

$$\rho_t = \frac{cov\,(X_t, Y_t)}{\sigma_{xt}\sigma_{yt}} = \frac{\frac{v}{v-2}\,\rho}{\frac{v}{v-2}} = \rho$$

The kurtosis of the resulting t distribution is computed as follows:[5]

$$\gamma_2 = \frac{6}{df-4}$$

Hence, I model the scenario of moderate excess of kurtosis ($\gamma_2 = 1$), using t distribution with 10 degrees of freedom.

The bivariate double exponential distribution $(X_{de}, Y_{de}) \sim DE(\sigma_x, \sigma_y, \rho)$(each marginal with excess of kurtosis 3) is generated by multiplying the original bivariate normal deviates (X^*, Y^*) (6) by the square-root of independent exponentially distributed variables:

$$\begin{pmatrix} X_{de} \\ X_{de} \end{pmatrix} = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} * \sqrt{\begin{pmatrix} a \\ b \end{pmatrix}} \quad \text{——(7)}$$

where a and b are the exponential variables with rate of $\lambda = 1$.[13] It can be shown that the Pearson correlation computed for the double exponential distribution generated by (7) equals to the Pearson correlation of the original bivariate normal data, indeed:

$$cov\,(X_{de}, Y_{de}) = \sigma_x \sigma_y \rho \lambda$$

$$V(X_{de}) = \frac{\sigma_x^2}{\lambda};\ V(Y_{de}) = \frac{\sigma_y^2}{\lambda}$$

The excess of kurtosis for the double exponential distribution equals to 3 regardless of λ. This could also be confirmed by direct computations using the closed form of moment generating function of double exponential distribution (derivations are not shown).

3.3 PARAMETERS OF THE SIMULATION STUDY

A simulation study is designed to assess the type I error rate and the relative statistical power of the permutation tests based on the Pearson's, Spearman's and Kendall's correlation coefficients. For my simulations, I consider sample sizes of 10, 20, 50 and 100, and the true Pearson correlations of 0, 0.2, 0.4, 0.6 and 0.8, which allows me to consider the scenarios where I expect to see substantial differences as well as the scenarios where the approaches are approximately equivalent. For estimation of type I error rate (scenario corresponding to ρ of 0), I generate 10,000 datasets and for estimation of statistical power (scenarios corresponding to non-zero ρ), I generate 1,000 datasets. For each of the generated datasets, the permutation test is implemented on 5,000 random permutation samples. The rejection rate (type I error rate or statistical power depending on the scenario) is estimated as the proportion of the simulated datasets for which I observe permutation p-value less than 0.05, i.e.:

$$R = \frac{\#\ of\ (P < 0.05)}{\#\ of\ simulation}$$

where P is the permutation p-value defined in equation (5) in the Permutation Test section.

For the considered distributions the scenarios where the true Pearson's correlation coefficient is zero correspond to the absence of any associations between two variables. Hence,

for these cases when the correlation is zero, the rejection rate corresponds to the type I error rate. The rejection rate is the statistical power of the test when the correlation is non-zero.

4.0 RESULTS

Tables 4, 5 and 7 summarize the estimates of rejection rates for the permutation tests based on the three correlation coefficients for various distributions. There is no substantial difference in type I error rates (columns with $\rho = 0$) and all estimates are close to nominal 0.05. As expected, the statistical power (columns with $\rho \neq 0$) increases as the true Pearson's correlation and sample size increase.

As expected, for bivariate normal distribution (Table 4), the Pearson product moment correlation consistently results a larger statistical power. Since the bivariate normal distribution is optimal for the Pearson's correlation coefficient, the increase of statistical power observed in Table 4 indicates the largest magnitude for the gains that can be observed in other distributions. Comparing with Kendall's tau, Spearman's correlation coefficient leads to a somewhat more powerful test. The dissimilarity of the statistical power values among the three correlation coefficients become less significant for moderate combinations of the true Pearson's correlation and sample size. The differences diminish as the statistical power approaches 1 (increasing sample size for given correlation, or increasing correlation for moderate sample size). The differences also diminish when the statistical power approaches 0 (decreasing sample size for given correlation, or decreasing correlation for given sample size). These trends also occur for the other of considered distributions. Although the Pearson's correlation coefficient results in a theoretically more powerful test for bivariate normal scenario, regardless of the true value of

correlation, the estimates of the statistical power could be smaller than the other procedures (e.g. $\rho = 0.2$, $n = 10$). These observations are caused purely by sampling error.

Table 5 demonstrates the relative performance of the Pearson product moment correlation coefficient for log-normal data (skewed distributions). As described in Section 3.2, the log-normal data is generated as an order-preserving transformation of the bivariate normal data used in Table 4. Hence the ranks of observations, and therefore the values of Spearman's and Kendall's correlation coefficient are exactly the same as for normal data.

For moderately skewed distribution ($\gamma_1 = 0.95$), the test based on the Pearson's correlation coefficient remains the most powerful with a gain in statistical power as high as 0.09. For a more severely skewed distribution ($\gamma_1 = 1.75$), the Pearson's correlation coefficient gradually loses its advantages (especially for larger sample sizes or large correlations). For most of the considered scenarios, it leads to a loss of statistical power with the maximum magnitude of 0.03. However, it can still lead to a substantial increase of statistical power in some cases (e.g. $\rho = 0.8$, $n = 10$).

Table 7 illustrates the relative performance of the correlation coefficients for distributions with excess of kurtosis. For distribution with moderate excess of kurtosis ($\gamma_2 = 1$), the Pearson product moment correlation consistently perform better than Spearman's and Kendall's correlation coefficients, leading to a gain in statistical power as high as 0.1. Kendall's tau, on the other hand, tends to perform slightly worse than Spearman's correlation coefficient. For the distribution with a severe excess of kurtosis ($\gamma_2 = 3$), the Pearson's correlation coefficient leads to almost a uniform loss in statistical power which can be as high as 0.08.

Table 4. Rejection rates for the bivariate normal distribution*

Distribution	Sample Size	Correlation Coefficient	ρ				
			Type I Error Rate	Statistical Power			
			0	0.2	0.4	0.6	0.8
Normal	10	r	0.0474	0.073	0.216	0.479	0.861
		r_s	0.0503	0.080	0.197	0.404	0.760
		τ	0.0472	0.076	0.181	0.401	0.752
	20	r	0.0478	0.126	0.412	0.832	0.996
		r_s	0.0481	0.121	0.372	0.772	0.991
		τ	0.0494	0.115	0.362	0.773	0.992
	50	r	0.0482	0.280	0.836	0.997	1
		r_s	0.0492	0.261	0.776	0.993	1
		τ	0.0490	0.261	0.778	0.995	1
	100	r	0.0525	0.516	0.987	1	1
		r_s	0.0529	0.469	0.978	1	1
		τ	0.0530	0.460	0.977	1	1

* No of simulation for sample size 10 is 10,000
* No of simulation for sample size 20, 50 and 50 is 1,000
* No of permutation for all sample size is 5,000

Table 5. Rejection rates for the skewed distributions*

Distribution	Sample Size	Correlation Coefficient	ρ				
			Type I Error Rate	Statistical Power			
			0	0.2	0.4	0.6	0.8
Log-normal γ_1 = 0.95	10	r	0.0491	0.081	0.220	0.460	0.850
		r_s	0.0503	0.080	0.197	0.404	0.760
		τ	0.0472	0.076	0.181	0.401	0.752
	20	r	0.0511	0.129	0.397	0.809	0.995
		r_s	0.0481	0.121	0.372	0.772	0.991
		τ	0.0494	0.115	0.362	0.773	0.992
	50	r	0.0480	0.265	0.800	0.997	1
		r_s	0.0492	0.261	0.776	0.993	1
		τ	0.0490	0.261	0.778	0.995	1
	100	r	0.0526	0.500	0.981	1	1
		r_s	0.0529	0.469	0.978	1	1
		τ	0.0530	0.460	0.977	1	1
Log-normal γ_1 = 1.75	10	r	0.0497	0.087	0.224	0.437	0.822
		r_s	0.0503	0.080	0.197	0.404	0.760
		τ	0.0472	0.076	0.181	0.401	0.752
	20	r	0.0511	0.129	0.377	0.760	0.991
		r_s	0.0481	0.121	0.372	0.772	0.991
		τ	0.0494	0.115	0.362	0.773	0.992
	50	r	0.0453	0.260	0.745	0.991	1
		r_s	0.0492	0.261	0.776	0.993	1
		τ	0.0490	0.261	0.778	0.995	1
	100	r	0.0513	0.457	0.958	1	1
		r_s	0.0529	0.469	0.978	1	1
		τ	0.0530	0.460	0.977	1	1

* No of simulation for sample size 10 is 10,000
* No of simulation for sample size 20, 50 and 50 is 1,000
* No of permutation for all sample size is 5,000

Table 6. True values of the Pearson's correlation coefficient ρ for log-normal data

		ρ*				
	$\sigma_x=\sigma_y$	0	0.2	0.4	0.6	0.8
ρ_{ln}#	0.3	0	0.19	0.39	0.59	0.79
	0.5	0	0.18	0.37	0.57	0.78

* Correlation for normal data
# Correlation for log-normal data

Table 7. Rejection rates for the distributions with excess kurtosis*

Distribution	Sample Size	Correlation Coefficient	ρ				
			Type I Error Rate	Statistical Power			
			0	0.2	0.4	0.6	0.8
T $\gamma_2 = 1$	10	r	0.0480	0.082	0.182	0.409	0.814
		r_s	0.0469	0.079	0.173	0.369	0.708
		τ	0.0465	0.072	0.157	0.338	0.701
	20	r	0.0500	0.132	0.400	0.792	0.989
		r_s	0.0467	0.120	0.381	0.754	0.977
		τ	0.0480	0.121	0.366	0.758	0.978
	50	r	0.0492	0.251	0.768	0.996	1
		r_s	0.0501	0.243	0.744	0.993	1
		τ	0.0506	0.240	0.735	0.993	1
	100	r	0.0482	0.484	0.978	1	1
		r_s	0.0488	0.451	0.973	1	1
		τ	0.0483	0.454	0.975	1	1
Double exponential $\gamma_2 = 3$	10	r	0.0448	0.073	0.149	0.319	0.600
		r_s	0.0494	0.081	0.161	0.327	0.594
		τ	0.0462	0.075	0.152	0.302	0.574
	20	r	0.0499	0.089	0.281	0.571	0.912
		r_s	0.0494	0.101	0.285	0.630	0.941
		τ	0.0512	0.106	0.278	0.621	0.930
	50	r	0.0495	0.219	0.618	0.952	1
		r_s	0.0488	0.220	0.703	0.971	1
		τ	0.0491	0.218	0.696	0.974	1
	100	r	0.0487	0.322	0.909	0.999	1
		r_s	0.0484	0.377	0.942	0.999	1
		τ	0.0484	0.381	0.941	1	1

* No of simulation for sample size 10 is 10,000
* No of simulation for sample size 20, 50 and 50 is 1,000
* No of permutation for all sample size is 5,000

5.0 CONCLUSION

I have demonstrated that using the permutation approach, the Pearson product moment correlation coefficient could be successfully used for analysis of continuous non-normally distributed data. The permutation test based on the Pearson's correlation coefficient, as well as the permutation tests based on Spearman's and Kendall's correlation coefficients, has type I error rate that is close to the nominal 0.05.

The use of the Pearson product moment correlation coefficient leads to an almost uniform gain in statistical power not only for the bivariate normal distribution, but also for some moderately non-normal distributions. Specifically, I observed the substantial advantages of the Pearson correlation coefficient for the log-normal distribution with moderate excess of skewness ($\gamma_1 = 0.95$) and t distribution with moderate excess of kurtosis ($\gamma_2 = 1$).

For distributions with severe departures from normality, the Pearson's correlation loses its advantages. For the severely skewed log-normal distribution ($\gamma_1 = 1.75$), the Pearson product moment correlation coefficient leads to a lower statistical power in more than half of the cases. For the double exponential distribution, which has a severe excess of kurtosis ($\gamma_2 = 3$), the Pearson's correlation coefficient results in a loss of statistical power in almost all of the scenarios I considered. However, even in those cases where the use of Pearson's correlation was disadvantageous, the maximum loss was less than the maximum possible gain in more "regular" distributions.

In conclusion, the permutation test based on the Pearson product moment correlation could offer a valuable advantage over Spearman's and Kendall's correlation coefficients in continuous non-normal distributions, for which the appropriateness of the Person's correlation could be questionable according to existing guidelines.

Thus, the sole fact of non-normality of the distribution should not be a sufficient reason for disregarding the use of the Pearson product moment correlation for continuous data.

6.0 DISCUSSION

This work provides some evidence of the advantages of the Pearson product moment correlation for the distributions under which its use is not advised by the current guidelines. The superiority of the Pearson's over Spearman's and Kendall's correlations stems from the fact that Pearson correlation better reflects the degree of concordance and discordance of pairs of observations for some types of distributions. Disadvantages of the Pearson product moment correlation seem to be mostly due to its known sensitivity to outliers.

Indeed, both increase of skewness and excess of kurtosis of the distributions of correlated variables increase the possibility of the outliers, and result in increasingly poorer performance of the Pearson's correlation (as compared with more outlier-insensitive Spearman's and Kendall's correlation). This is more evident for large sample sizes where the probability of obtaining datasets with the outliers is higher.

However, for moderately non-normal distributions, where the outliers are possible but not as frequent, the Pearson product moment correlation can still lead to a substantial gain in statistical power. Thus, the guidelines for the appropriateness of the Pearson's correlation for continuous normal data should be primarily based on the evidence of outliers in the data, rather than on shape of the empirical or theoretical distribution.

The presented investigation is somewhat preliminary. It considers two specific types of departures from normality independently. Furthermore, the conclusions relate only to the

permutation tests and do not immediately generalize to the statistical tests standard for the considered correlation coefficients.[19] Finally here I do not investigate any of the approaches for determining whether a given dataset contains any outliers. Future investigations could benefit from considering the distributions that demonstrate excess of skewness and kurtosis simultaneously, as well as a standard test for the Pearson product moment correlation.

The investigation of the relative statistical power summarized in this work has direct practical implications. Indeed, the gain in statistical power offered by the Pearson's correlation coefficient directly translates into the increased possibility of obtaining conclusive results of the data analysis and reduction of the sample size for future studies. In regard to the analysis of the already collected datasets, the higher statistical power implies that, the use of the Pearson's correlation coefficient could result in a statistically significant result when results based on other correlation coefficients are insignificant, hence inconclusive. For the study design purposes, the use of the Pearson's correlation coefficient could lead to a smaller sample size estimate, thereby saving the resources for conducting a future study. However, the sample size gain could be only crudely estimated from the presented results, and more precise estimates require a different simulation study which could be considered in the future investigations.

APPENDIX

SAS CODE FOR PERMUTATION TEST AND SIMULATION STUDY

1. THE GENERATION OF SAMPLING DISTRIBUTIONS

```
/* Simulations */
%macro set(n,n_sim,rho,output);
proc iml;

      x=j(&n,1,0);
      y=j(&n,1,0);
      z=j(&n,1,0);

      seed=54321;

      mu={0 0};

      sx=1;
      sy=sx;

      sig=((sx**2)||(&rho*sx*sy))//((&rho*sx*sy)||(sy**2));
      sigma=root(sig);

      do sim=1 to &n_sim;

            call rannor(seed,x);
            call rannor(seed,y);

            xy=x||y;

            data=repeat(mu,&n,1)+xy*sigma;

            datax=data[,1];
            datay=data[,2];


            /* Pearson */
```

```
            pearsono=corr(data);

            r=pearsono[1,2]; /* Pearson’s r */

            po=po//r; /* column vector Pearson’s r */

            po2=po2//(r>0); /* column vector of Pearson’s r > 0 */


            /* Spearman's */

            data2=(rank(data[,1]))||(rank(data[,2])); /* ranked data */

            spearmano=corr(data2);

            rs=spearmano[1,2]; /* Spearman's rho */

            so=so//rs; /* column vector of Spearman's rho */

            so2=so2//(rs>0); /* column vector of Spearman's rho > 0 */


            /* Kendall's */

            xo1=repeat(datax,1,&n);
            xo_to_xo=(xo1<xo1`)+((xo1=xo1`)*0.5);
            xo2=(xo_to_xo*2)-1;

            yo1=repeat(datay,1,&n);
            yo_to_yo=(yo1<yo1`)+((yo1=yo1`)*0.5);
            yo2=(yo_to_yo*2)-1;

            sumo=(xo2#yo2)[+];

            t=sumo/(&n*(&n-1)); /* Kendall's tau */

            ko=ko//t; /* column vector of Kendall's tau */

            ko2=ko2//(t>0);   /* column vector of Kendall's tau > 0 */


      end;

sump=po2[+];

sums=so2[+];

sumk=ko2[+];

print sump sums sumk;

create &output var{po so ko};
append;


quit;
```

```
%mend;

/* Coefficients that greater than zero */
%set(10,10000,0.4,corr1)
%set(20,10000,0.4,corr2)
%set(50,10000,0.4,corr3)
%set(100,10000,0.4,corr4)


%macro plot(corr,var1,var2,var3);
/* Constructing histogram for Pearson product moment correlation */

proc univariate data=&corr noprint;
      histogram &var1 / midpoints=-1.2 to 1.2 by 0.05;
      label po ='r';
      title "Pearson product moment correlation coefficient";
run;

/* Constructing histogram for Spearman's rank-order correlation*/

proc univariate data=&corr noprint;
      histogram &var2 / midpoints=-1.2 to 1.2 by 0.05;
      label so ='r';
      title "Spearman's rank-order correlation coefficient";
run;

/* Constructing histogram for Kendall's tau correlation*/

proc univariate data=&corr noprint;
      histogram &var3 / midpoints=-1.2 to 1.2 by 0.05;
      label ko ='tau';
      title "Kendall's tau correlation coefficient";
run;

%mend;

/* Correlation */
%plot(corr1,po,so,ko)
%plot(corr2,po,so,ko)


/* True values of Spearman's & Kendall's correlation coefficients */
%macro set(n,n_sim,rho);
proc iml;

      x=j(&n,1,0);
      y=j(&n,1,0);
      z=j(&n,1,0);

      seed=54321;

      mu={0 0};

      sx=1;
      sy=sx;
```

```
	sig=((sx**2)||(&rho*sx*sy))//((&rho*sx*sy)||(sy**2));
	sigma=root(sig);

	do sim=1 to &n_sim;

		call rannor(seed,x);
		call rannor(seed,y);

		xy=x||y;

		data=repeat(mu,&n,1)+xy*sigma;

		datax=data[,1];
		datay=data[,2];


		/* Spearman's */

		data2=(rank(data[,1]))||(rank(data[,2])); /* ranked data */

		spearmano=corr(data2);

		rs=spearmano[1,2]; /* Spearman's rho */

		so=so//rs; /* column vector of Spearman's rho */


		/* Kendall's */

		xo1=repeat(datax,1,&n);
		xo_to_xo=(xo1<xo1`)+((xo1=xo1`)*0.5);
		xo2=(xo_to_xo*2)-1;

		yo1=repeat(datay,1,&n);
		yo_to_yo=(yo1<yo1`)+((yo1=yo1`)*0.5);
		yo2=(yo_to_yo*2)-1;

		sumo=(xo2#yo2)[+];

		t=sumo/(&n*(&n-1)); /* Kendall's tau */

		ko=ko//t; /* column vector of Kendall's tau */


	end;

ts=so[:]; /* True value of Spearman's correlation coefficient */

tk=ko[:]; /* True value of Spearman's correlation coefficient */

print ts tk;


quit;

%mend;
```

```
%set(500,1000,0)
%set(500,1000,0.2)
%set(500,10000,0.4)
%set(500,1000,0.6)
%set(500,1000,0.8)
```

2. THE GENERATION OF NORMAL DISTRIBUTED DATA

```
/* Simulations & Permutations */
%macro set(n,rho,n_sim,n_perm);
proc iml;

      x=j(&n,1,0);
      y=j(&n,1,0);
      z=j(&n,1,0);

      seed=54321;

      mu={0 0};

      sx=1;
      sy=sx;

      sig=((sx**2)||(&rho*sx*sy))//((&rho*sx*sy)||(sy**2));
      sigma=root(sig);

      do sim=1 to &n_sim;

            call rannor(seed,x);
            call rannor(seed,y);

            xy=x||y;

            data=repeat(mu,&n,1)+xy*sigma;

            datax=data[,1];
            datay=data[,2];


            /* Pearson's */

            pearsono=corr(data);

            r=pearsono[1,2]; /* original Pearson's r */

            po=po//r; /* column vector of original Pearson's r */


            /* Spearman's */
```

```
d=(rank(datax))||(rank(datay)); /* ranked data */

spearmano=corr(d);

rs=spearmano[1,2]; /* original Spearman's rho */

so=so//rs; /* column vector of original Spearman's rho */


/* Kendall's */

xo1=repeat(datax,1,&n);
xo_to_xo=(xo1<xo1`)+((xo1=xo1`)*0.5);
xo2=(xo_to_xo*2)-1;

yo1=repeat(datay,1,&n);
yo_to_yo=(yo1<yo1`)+((yo1=yo1`)*0.5);
yo2=(yo_to_yo*2)-1;

sumo=(xo2#yo2)[+];

t=sumo/(&n*(&n-1));     /* original Kendall's tau */

ko=ko//t; /* column vector of original Kendall's tau */


/* Permutations */

do perm=1 to &n_perm;

      call ranuni(seed,z);

      temp=data[,2]||z;

      call sort(temp,{2});

      data2=data[,1]||temp[,1]; /* permuted data */

      data2x=data2[,1];
      data2y=data2[,2];


      /* Pearson's */

      pearson=corr(data2);

      r2=pearson[1,2]; /* permuted Pearson's r */

      p=p//r2; /* column vector of permuted Pearson's r */


      /* Spearman's */

      d2=(rank(data2x))||(rank(data2y)); /* ranked data */

      spearman=corr(d2);
```

```
        rs2=spearman[1,2]; /* permuted Spearman's rho */

        s=s//rs2; /* column vector of permuted Spearman's rho */


        /* Kendall's */

        x1=repeat(data2x,1,&n);
        x_to_x=(x1<x1`)+((x1=x1`)*0.5);
        x2=(x_to_x*2)-1;

        y1=repeat(data2y,1,&n);
        y_to_y=(y1<y1`)+((y1=y1`)*0.5);
        y2=(y_to_y*2)-1;

        sum=(x2#y2)[+];

        t2=sum/(&n*(&n-1));     /* permuted Kendall's tau */

        k=k//t2; /* column vector of permuted Kendall's tau */


end;


/* Mid Point P-Value */

/* Pearson's */

c1p=((abs(p))>(abs(r)));
c2p=((abs(p))=(abs(r)));

pvalp=(c1p[:])+((c2p[+])/(2*&n_perm)); /* Pearson's p-values */

rejectp=rejectp//(pvalp<0.05); /* Pearson's rejection */


/* Spearman's */

c1s=((abs(s))>(abs(rs)));
c2s=((abs(s))=(abs(rs)));

pvals=(c1s[:])+((c2s[+])/(2*&n_perm)); /* Spearman's p-values */

rejects=rejects//(pvals<0.05); /* Spearman's rejection */


/* Kendall's */

c1k=((abs(k))>(abs(t)));
c2k=((abs(k))=(abs(t)));

pvalk=(c1k[:])+((c2k[+])/(2*&n_perm)); /* Kendall's p-values */

rejectk=rejectk//(pvalk<0.05); /* Kendall's rejection */
```

```
            free p s k;

      end;

rratep=rejectp[:]; /* Pearson's rejection rate */

rrates=rejects[:]; /* Spearman's rejection rate */

rratek=rejectk[:]; /* Kendall's rejection rate */


print rratep rrates rratek;


quit;

%mend;

%set(10,0,10000,5000)
%set(10,0.2,1000,5000)
%set(10,0.4,1000,5000)
%set(10,0.6,1000,5000)
%set(10,0.8,1000,5000)

%set(20,0,10000,5000)
%set(20,0.2,1000,5000)
%set(20,0.4,1000,5000)
%set(20,0.6,1000,5000)
%set(20,0.8,1000,5000)

%set(50,0,10000,5000)
%set(50,0.2,1000,5000)
%set(50,0.4,1000,5000)
%set(50,0.6,1000,5000)
%set(50,0.8,1000,5000)

%set(100,0,10000,5000)
%set(100,0.2,1000,5000)
%set(100,0.4,1000,5000)
%set(100,0.6,1000,5000)
%set(100,0.8,1000,5000)
```

3. THE GENERATION OF LOG-NORMAL DISTRIBUTED DATA

```
/* Simulations & Permutations */
%macro set(n,rho,n_sim,n_perm);
proc iml;
```

```
      x=j(&n,1,0);
      y=j(&n,1,0);
      z=j(&n,1,0);

      seed=54321;

      mu={0 0};

      sx=0.3;
      sy=sx;

      sig=((sx**2)||(&rho*sx*sy))//((&rho*sx*sy)||(sy**2));
      sigma=root(sig);

      do sim=1 to &n_sim;

            call rannor(seed,x);
            call rannor(seed,y);

            xy=x||y;

            data=exp(repeat(mu,&n,1)+xy*sigma);

            datax=data[,1];
            datay=data[,2]

/* Setting sx=sy=0.5 for second scenario */


%mend;
```

4. THE GENERATION OF T DISTRIBUTED DATA

```
/* Simulations & Permutations */
%macro set(n,rho,n_sim,n_perm);
proc iml;

      x=j(&n,1,0);
      y=j(&n,1,0);
      u=j(&n,1,0);
      v=j(&n,1,0);
      z=j(&n,1,0);


      seed=54321;

      mu={0 0};

      sx=1;
```

```
      sy=sx;

      df=10;

      sig=((sx**2)||(&rho*sx*sy))//((&rho*sx*sy)||(sy**2));
      sigma=root(sig);

      do sim=1 to &n_sim;

            call rannor(seed,x);
            call rannor(seed,y);
            call rangam(seed,(df/2),u);
            call rangam(seed,(df/2),v);

            xy=x||y;

            w=repeat(mu,&n,1)+xy*sigma; /* Bivariate normal variable w */

            c=2*(u||v); /* Chi-squared variable c */

            data=w#(sqrt(df/c));

            datax=data[,1];
            datay=data[,2];


%mend;
```

5. THE GENERATION OF DOUBLE EXPONENTIAL DISTRIBUTED DATA

```
/* Simulations & Permutations */
%macro set(n,rho,n_sim,n_perm);
proc iml;

      x=j(&n,1,0);
      y=j(&n,1,0);
      a=j(&n,1,0);
      b=j(&n,1,0);
      z=j(&n,1,0);


      seed=54321;

      mu={0 0};

      sx=1;
      sy=1;

      sig=((sx**2)||(&rho*sx*sy))//((&rho*sx*sy)||(sy**2));
      sigma=root(sig);
```

```
	do sim=1 to &n_sim;

		call rannor(seed,x);
		call rannor(seed,y);
		call ranexp(seed,a);
		call ranexp(seed,b);

		xy=x||y;

		w=repeat(mu,&n,1)+xy*sigma; /* Bivariate normal variable w */

		e=a||b; /* Exponential variable e */

		data=(sqrt(e))#w;

		datax=data[,1];
		datay=data[,2];


%mend;
```

BIBLIOGRAPHY

1. Abdullah MB. On a Robust Correlation Coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*. 1990;39:455-460.

2. Balakrishnan N, Lai CD. *Continuous Bivariate Distributions*. 2nd ed. New York: Springer; 2009.

3. Chen PY, Popovich PM. *Correlation: Parametric and Nonparametric Measures*. Thousand Oaks, CA: Sage Publications, Inc.; 2002.

4. Corty EW. *Using and Interpreting Statistics: A Practical Text for the Health, Behavioral, and Social Sciences*. St. Louis, MO: Mosby, Inc; 2007.

5. DeCarlo LT. On the Meaning and Use of Kurtosis. *Psychological Methods*. 1997;2(3):292-307.

6. Ernst MD. Permutation Methods: A Basis for Exact Inference. *Statistical Science.* 2004;19:676-685.

7. Gibbons JD. *Nonparametric Measures of Association.* Newbury, CA: Sage Publications, Inc.; 1993.

8. Johnson NL, Kotz S & Balakrishnan N. *Continuous Univariate Distributions: Volume 1*. 2nd ed. New York: John Wiley & Sons, Inc; 1994.

9. Kendall MG. A New Measure of Rank Correlation. *Biometrika*. 1938;30:81-93.

10. Kendall MG. *Rank Correlation Methods*. 3rd ed. New York: Hafner Publishing Company; 1962.

11. Khamis H. Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography*. 2008;24:155-162.

12. Kotz S, Balakrishnan N, Johnson NL. *Continuous Multivariate Distributions: Volume 1: Models and Applications*. 2nd ed. New York: John Wiley & Sons, Inc; 2000.

13. Kotz S, Kozubowski TJ, Podgórski K. *The Laplace Distribution and Generalizations: a revisit with applications to communications, economics, engineering, and finance*. New York: Birkhäuser Boston, c/o Springer-Verlag New York, Inc; 2001.

14. Kotz S, Nadarajah S. *Multivariate t Distributions and Their Applications*. New York: Cambridge University Press; 2004.

15. Lewis-Beck MS. *Data Analysis: An Introduction.* Thousand Oaks, CA: Sage Publications, Inc.; 1995.

16. Lieberson S. Limitations in the Application of Non-Parametric Coefficients of Correlation. *American Sociological Review*. 1964;29:744-746.

17. Liebetrau AM. *Measures of Association.* Beverly Hills and London: Sage Publications, Inc.; 1976.

18. McKillup S. *Statistics Explained: An Introductory Guide for Life Sciences.* Cambridge, UK: Cambridge University Press; 2005.

19. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures.* 4th ed. Boca Raton, FL: Chapman & Hall/CRC; 2007.

20. Siegel S. Nonparametric Statistics. *The American Statistician*. 1957;11:13-19.

21. Sommer R, Sommer B. *A Practical Guide to Behavioral Research: Tools and Techniques*. 5th ed. New York: Oxford University Press, Inc; 2002.

22. Sprent P, Smeeton NC. *Applied Nonparametic Statistical Methods*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC; 2007.

23. Taylor JMG. Kendall's and Spearman's Correlation Coefficients in the Presence of a Blocking Variable. *Biometrics*. 1987;43:409-416.